DOCUMENT RESUME

ED 440 987 TM 030 783

AUTHOR Ediger, Marlow
TITLE How Should Teachers Be Assessed?
PUB DATE 2000-02-00
NOTE 10p.
PUB TYPE Opinion Papers (120)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Criterion Referenced Tests; *Evaluation Methods; *Portfolio Assessment; Scoring Rubrics; Teacher Effectiveness; *Teacher Evaluation; Test Results; Test Use

ABSTRACT

Achieving a fair and objective evaluation of teacher performance is not easy. Student test results have been used to assess teacher performance, but this approach has many weaknesses, especially because students differ from school to school and standardized tests are open to questions about their fairness. Criterion-referenced tests have been developed to counteract some weaknesses apparent in norm-referenced standardized tests, but even these tests have weaknesses. For one thing, student performance is measured on only one occasion. Other criticisms may be leveled for the out-of-context nature of criterion-referenced tests. Portfolio use, however, stresses a philosophy of contextualism in appraising teacher effectiveness. Assessment through portfolios is not isolated and does not stress a numerical result unless rubrics are used. Even then, rubric ratings are open to human interpretation. Portfolios and their contents need to be aligned with stated objectives to be fair assessments, but their use can help gauge student learning and teacher effectiveness. (SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

ED 440 987

How Should Teachers Be Assessed?

Marlow Ediger

TM030783

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION
CENTER (ERIC)
This document has been reproduced as received from the person or organization originating it.
Minor changes have been made to improve reproduction quality.
Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

Marlow Ediger

1

BEST COPY AVAILABLE

HOW SHOULD TEACHERS BE ASSESSED?

Much is written and discussed pertaining to teacher accountability. All people need to be accountable to achieve objectives in society. When the writer came to teach at Truman State University in 1962, then Northeast Missouri State Teacher's College, a consultant spoke to the division of Education on being accountable, well ahead of the time when this concept became a household word to teachers and administrators. He mentioned that as a teacher anyone could come into his classroom to observe and notice the quality of his teaching. Nothing was said there about documenting teaching quality. When thinking back about the consultant's ideas, there appears to be much merit in what he was saying. First hand then, an observer could come into the classroom to evaluate, comment after the class visit, and assess teaching quality.

There needs to be a just way of assessing teaching performance. At the present time, there are many variables which hinder some kind of agreed upon objective and fair way of assessing teacher progress. By discussing the pros and cons of each procedure, perhaps, a synthesis may be reached. Different schools of thought in assessing teaching performance have unique inherent philosophies. These need analyization in order to come up with a synthesis in moving from <u>what is</u> to <u>what should be</u>.

The Measurement Movement

The measurement movement has been with educators since the beginning of the twentieth century. E. L. Thorndike (1874-1949) expressed selected basic measurement ideas such as the following in the early 1900s:

1. Whatever exists, exists in some amount.
2. If it exists in some amount, it can be measured.

Thorndike and his associates were busy with developing diverse measurement instruments to ascertain numerically what students had learned and acquired in different subject matter areas (See Thorndike, 1918). The steady growth in number of standardized tests resulted. With standardization, each student in a class

1. took the same test.
2. had the same time limits for taking that test.
3. was given the same directions for taking the test.
4. had tests scored using the same answer key.
5. had test results compared with the same norm group that the test was standardized on.

Numerical results were and are provided to indicate how well a student did on the standardized test. Percentile rating are generally provided. However, standard deviations above and below the mean, quartile deviations, stanine scores, and grade equivalents also may be provided to reveal student performance. With a numeral, such as a percentile, being given to indicate student performance, quick methods are used to report student achievement to parents and the lay public. Thus, a single numeral "tells it all."

Test results from students have been used to reveal the quality of teaching of a specific teacher. However, there are many weaknesses here:

1. the playing field of students from school to school is anything but level. Students from suburbia have always achieved much higher on standardized tests as compared to urban and rural learners. They have had much better opportunities to learn n the preschool years as well as after school. Selected educators believe that standardized tests measure socio-economic levels rather than academic learnings.

2. standardized tests have their many weaknesses. Thus, with standardized means of assessing, "one size fits all." And yet, students differ from each other in many ways, such as abilities possessed, quickness in responding to test items, motivation, and purpose in learning.

3. standardized tests claim to measure objectively. However, there is little objectivity when subject matter is selected to be in a test. Human beings make these subjective decisions. Certain included subject matter will be more familiar to selected students as compared to others due to the region the school is in. Selected units of study taught in a school or school system may appear, in part, on a standardized test. Other schools may have taught selected subject matter not appearing on the test. For all to have had equal opportunities to learn subject matter contained in a test is impossible.

4. weaknesses exist also in the many assumptions that standardized tests operate on. Thus, a question may arise on how valid a test is even though the correlation figures presented in the manual is high. To truly be valid, the learning activities in any school need to be aligned with the objectives of the test. And yet, standardized tests have no objectives for teachers to use as benchmarks in teaching students in the classroom.

5. standardized tests have been criticized for stressing reliability to the minimizing of validity concepts. Thus, standardized tests in many cases do have a difficult time of relating test items thereon to what is being taught in the school setting. Students in the class setting will be tested with the use of standardized tests at selected intervals. Consistency of results in student testing is easier to obtain when using test/retest, alternate forms, or split hall reliability, as compared to

having demonstrated validity in a test.

It becomes difficult to make a case in holding teachers accountable for student results after the latter has taken the involved standardized test (See Ediger, 1994, 169-174).

Teacher Assessment and Criterion Referenced Tests (CRTs)

To remedy a deficiency, CRTs were developed, usually under the auspices of state departments of education. CRTs do have accompanying objectives for teachers to use as benchmarks for teaching students. Test items on the CRT tend to be more valid as compared to standardized tests in that they relate directly to the stated objectives used by the teacher in teaching. However, sometimes, the objectives are too open ended, making it difficult for the test item to truly reflect and measure what is in the stated objective(s).

Companies developing and selling standardized tests have more money available than do state departments of education to do research on the quality of their tests. Generally, multiple choice test items are used on standardized and CRTs. Weak test items need to be taken out in pilot studies conducted. Thus with a printout of student results in a pilot study, the evaluator of these tests may view the quality of each item by observing test item analysis therein. If all students, for example, were correct in responding to a test item, the chances are that test item lacks sophistication to notice student achievement. Or, if all students responded incorrect to a test item, the chances are that multiple choice item needs to be evaluated in terms of clarity in writing.

A further problem in using tests to measure student achievement and determining related teacher effectiveness in teaching is to decide upon how high the standards should be for learner achievement. In an era when "high standards" for student achievement are being emphasized, a problem arises as to how high should the hurdle be. Test items can be written on a level whereby each taker of the test can be successful in responding. The converse may also be said in that a test can be written at a highly complex level in which all or most would be unsuccessful in test results from having taken the test. The writer has mentioned this frequently in his Methods of Research class by giving examples in class for graduate students to respond to, such as what is 1 +1, 1+2, and 1+3 whereby all are correct 100 % of the time, unless human error comes in. Then too, test items may be written at a very complex level so that all/nearly all would fail, especially if validity is lacking in testing. With multiple choice test items, there is a one in four chance of guessing correctly as to which is the correct answer. If one or two of the responses in a multiple choice test item are ridiculous, the correct answer is much like a true/false item in that either one or the

other answer is correct

With testing in a paper/pencil situation, realism and practicality in the situation is a problem. For example, individuals at the work place are not tested to show proficiency. Rather they do the work required to demonstrate quality and productivity. Perhaps, in assessing achievement in the school setting, more opportunities should be provided students whereby they indicate in a practical situation what has been achieved. Thus, in units on Citizenship and Ethics, can the student truly show quality traits of character when interacting with others? Test results may not stress this interaction which is at the heart of citizenship and ethics education (See Ediger, 1995, Chapter Seven).

Multiple Intelligences Theory (See Gardner, 1993) indicate the need for students to be able to show what has been learned through diverse procedures, not paper/pencil tests only. These intelligences include verbal, kinesthetic, artistic, interpersonal, intrapersonal, musical, scientific, and mathematical. For example, a student may show what has been learned best through music when indicating learnings acquired from studying history. Many songs have been set forth in music that stress content in history. The point is that students should use the intelligence(s) possessed to show content and skills learned and not through paper/pencil testing only or largely. Gardner (2000) when asked in an interview why he criticized the facts based, standardized test approach in k-12 education responded with the following:

Facts are just bits and pieces of knowledge. They acquire meaning only when combined into significant patterns. Facts alone are like Christmas tree ornaments without a tree. Standardized tests that look at how many facts students know, as opposed to what they understand, force teachers to present these unconnected ideas. Then learning comes about choosing the correct answer on a multiple choice test. I think students should focus on a limited number of important topics, explore them in depth, and come to understand them well. Assessment ought to focus upon the kinds of things we want students to understand, and give kids a chance to perform their understanding.

Testing and measurement procedures to assess teaching performance then have their limitations with the following in criterion referenced testing:

1. one test is to show learner achievement even though given once a year at the most.
2. a single numeral here, such as a percentile, is to indicate student achievement over time.
3. isolated information provides the basis for testing with the use of multiple choice test items.
4. little, if any, feedback from student test results is provided to

teachers for diagnostic and remedial work.

5. a numeral is provided to parents to show their child's progress.

6. report cards may be issued in a state to compare test results in contrasting one school or school system with another. The playing field here is not level at all in comparing school achievement results involving suburbia, urban, and rural school achievement. Suburban teachers definitely have an advantage since their students are bound to achieve significantly higher than those from urban and rural schools.

7. CRTS, if not properly properly field tested, might have a large Standard Error of Measurement (SE meas). A large number of weaknesses may then be inherent in the involved CRT. CRTs need to have high reliability with alternative forms, split half, and/or test retest reliability.

8. openended objectives for teachers to use as benchmarks in teaching may not be precise enough to choose aligned learning activities. Validity then becomes weakened since the test items on the CRT should relate directly to the stated objectives.

9. fragmented leanings are generally measured in CRTs with its multiple choice test items.

10. testing does not occur in context within an ongoing lesson or unit of study. The test items are prepared by those outside of the local classroom (See Ediger, 1995, ERIC # ED 386319).

Accountability has been a key concept with the use of CRTs to gauge teacher effectiveness in terms of how much students have learned based on test results. There are numerous weaknesses here based on the above enumerated items.

Portfolios and Teacher Accountability

Portfolio use stresses a philosophy of contextualism when appraising teacher effectiveness in the classroom. Contextualism stresses assessing within ongoing lessons and units of study in terms of student learning. The assessing is not done by outsiders, removed from the local classroom. It does not deal with isolated bits of information that is observed on most standardized and CRTs. It does not stress a numerical result unless rubrics are used. However, even though the intent is to be objective, rubric ratings and their respective definitions tend to be quite openended and subject to human interpretation. Any product/process of a learner assessed in terms of rubric criteria, such as on a five point scale, will face problems of interrater/interscorer reliability.

Portfolios and their contents need to be aligned with stated objectives. Definite benchmarks for teaching and learning need to be in evidence. Anarchy in teaching should not be inherent and must be

avoided. To gauge student learning and teacher effectiveness, what might go into a portfolio?

1. written reports, outlines, summaries, poems, stories, plays, and expository information.
2. art products as they relate to ongoing lessons and units of study.
3. snapshots of construction work, dioramas, and friezes.
4. a video tape showing student participation in collaborative endeavors.
5. self evaluation results from the involved student.
6. classroom test scores based on specific units of study.
7. teacher evaluation of the student's progress.
8. diagrams and drawings illustrating ideas learned in subject matter.
9. cassette recordings of oral presentations.
10. computer printouts of completed work related directly to the local curriculum

A portfolio should not be too voluminous since they should be assessed by at least two professionals. Nor should they be too slender and not provide appropriate scope pertaining to what a student has achieved. What are selected drawbacks pertaining to portfolio use to assess teacher effectiveness?

1. they will tend to lack validity, in the traditional sense of testing philosophy, since the portfolio is quite openended in terms of relating to the stated objectives. The objectives too will tend to be more openended as compared to mensurable stated ends generally used for CRTs.
2. reliability will not be as precise and specific as compared to CRTS and standardized tests. Why? Interrater and Interscorer reliability may vary considerably from one assessor to the next.
3. if the services of paid assessors are used, the cost of assessing may be quite great. Machine scoring is impossible, presently, of portfolios.
4. much subjectivity is involved in students' determining what is to go into a portfolio. However, is it any more subjective as compared to test writers of standardized and CRTs ascertaining which content should be measured to indicate student achievement?
5. difficulties involved in parents and the lay public attaching meaning to voluminous portfolio contents to ascertain teacher effectiveness. A single numeral from a standardized or CRT is much easier to comprehend due to its simplicity (See Ediger, 1997, Chapter Five).

Additional Comments

There are numerous additional approaches that have been used to

assess teacher accountability. Teacher tests on the state level have been used to assess achievement in knowledge pertaining to teaching. If a teacher fails the state mandated test, he/she can lose credentials for teaching based on a single test's results. Generally, the test can be taken over by the teacher, if failure occurred on the initial test taken. Usually, state mandated tests for teachers attempt to show the defectiveness of undergraduate preparation programs for teaching. A single test is to reveal more than an entire undergraduate degree program from an approved teacher education degree institution.

A second procedure emphasizes bankruptcy laws in education. Thus, if an entire school has students that do not measure up to a selected standard, the school or school system may be taken over by the state and new administrators appointed and inservice education for teachers become a top priority. The entire school is then held accountable, not an individual teacher.

Third, workshops, faculty meetings, and clinical supervision have been used to upgrade teaching skills. A major problem here pertains to who should determine which the best procedures are to assist teachers to help students achieve as optimally as possible. For example, the philosophies of testing and measuring with standardized tests/CRTS is considerably different as compared to contextualism and portfolio use. Or, behaviorism with its measurably stated objectives, predetermined, is quite different from humanism with its student centered procedures of instruction.

Fourth, there are many issues involved in testing in and of itself, including a lack of agreement on the use of high stakes testing. Failing in a high stakes test may indeed be devastating to a student, such as not being able to graduate from high school. The involved tests used may not measure that accurately as selected educators may assume.

Fifth, the entire area of assessment is open for debate. Students are not like engines with standardized parts. They possess a feeling dimension, as human beings, which is definitely subjective. Each person has different feelings about happenings, events, goals in life, aims, and purposes. To standardize feelings violates a basic dimension of the human being and that is the subjective factor.

In closing, there are numerous articles written about what makes for effective schools. These need careful consideration. Cawalti (2000), former Executive Director of the Association for Supervision and Curriculum Development, lists the following benchmarks for quality schools:

1. a highly committed faculty.
2. strong leadership from the principal.
3. extensive time for reading.
4. extending time spent on task.
5. incentives and recognition.

6. a preassessment program and students practicing on what will be tested on involving the state mandated test.

References

Cawalti, Gordon (2000), "Portrait of a Benchmark School." Educational Leadership, 57 (5), 42-44.

Ediger, Marlow (2000), "Measurement and Evaluation," Studies in Educational Evaluation, 20 (2), 169-174.

Ediger, Marlow (1995), Philosophy in Curriculum Development. Kirksville, Missouri: Simpson Publishing Company, Chapter Seven.

Ediger, Marlow (1995), "To Every Action, There is an Opposite and Equal Reaction," Resources in Education, ERIC #ED 386319.

Ediger, Marlow (1997), The Modern Elementary School. Kirksville, Missouri: Simpson Publishing Company, Chapter Five.

Gardner, Howard (1993), Multiple Intelligences: Theory Into Practice. New York: Basic Books.

Gardner, Howard (2000), "Making the Most of Minds," Education Digest, 65 (6), 5.

Thorndike, E. L., "The Nature, Purpose, and General Methods of Measurement of Educational Products," 17th Yearbook, National Society or the Study of Education, Part Two.